

## ABSTRACT

A method for reducing memory latency in a multi-node architecture. In one embodiment, a speculative read request is issued to a home node before results of a cache coherence protocol are determined. The home node initiates a read to memory to complete the speculative read request. Results of a cache coherence protocol may be determined by a coherence agent to resolve cache coherency after the speculative read request is issued.